On RNA-RNA interaction structures of fixed topological genus

Benjamin M.M. Fu[a], Hillary S.W. Han[a], Christian M. Reidys[a,*]

[a] *Department of Mathematics and Computer science, University of Southern Denmark Campusvej 55, DK-5230 Odense M, Denmark*

Abstract

Interacting RNA complexes are studied via bicellular maps using a filtration via their topological genus. Our main result is a new bijection for RNA-RNA interaction structures and linear time uniform sampling algorithm for RNA complexes of fixed topological genus. The bijection allows to either reduce the topological genus of a bicellular map directly, or to lose connectivity by decomposing the complex into a pair of single stranded RNA structures. Our main result is proved bijectively. It provides an explicit algorithm of how to rewire the corresponding complexes and an unambiguous decomposition grammar. Using the concept of genus induction, we construct bicellular maps of fixed topological genus g uniformly in linear time. We present various statistics on these topological RNA complexes and compare our findings with biological complexes. Furthermore we show how to construct loop-energy based complexes using our decomposition grammar.

Keywords: RNA interaction structure, bicellular map, topological genus, genus induction, uniform generation, sampling

2010 MSC: 05A19, 92E10

*Corresponding author

Email addresses: benjaminfmm@imada.sdu.dk (Benjamin M.M. Fu), hillary@imada.sdu.dk (Hillary S.W. Han), duck@santafe.edu (Christian M. Reidys)

1. Introduction

RNA-RNA interactions constitute one of the fundamental mechanisms of cellular regulation. We find such interactions in a variety of contexts: small RNAs binding a larger (m)RNA target including: the regulation of translation in both prokaryotes [1] and eukaryotes [2, 3], the targeting of chemical modifications [4], insertion editing [5] and transcriptional control [6].

A salient feature is the formation of RNA-RNA interaction structures that are far more complex than simple sense-antisense interactions. This is observed for a vast variety of RNA classes including miRNAs, siRNAs, snRNAs, gRNAs, and snoRNAs. Thus deeper understanding of RNA-RNA interactions in terms of the thermodynamics of binding and in its structural consequences is a necessary prerequisite to understanding RNA-based regulation mechanisms.

An RNA molecule is a linearly oriented sequence of four types of nucleotides, namely, **A**, **U**, **C**, and **G**. This sequence is endowed with a well-defined orientation from the $5'$- to the $3'$-end and referred to as the backbone. Each nucleotide can form a base pair by interacting with at most one other nucleotide by establishing hydrogen bonds. Here we restrict ourselves to Watson-Crick base pairs **GC** and **AU** as well as the wobble base pairs **GU**. In the following, base triples as well as other types of more complex interactions are neglected.

RNA structures can be presented as diagrams by drawing the backbone horizontally and all base pairs as arcs in the upper half-plane; see Figure 1. This set of arcs provides our coarse-grained RNA structure in particular ignoring any spatial embedding or geometry of the molecule beyond its base pairs.

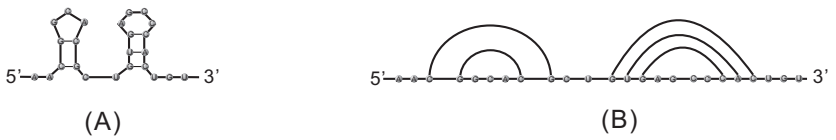

Figure 1: (A) An RNA secondary structure and (B) its diagram representation.

As a result, specific classes of base pairs translate into distinct structure

categories, the most prominent of which are secondary structures [7, 8, 9, 10]. Represented as diagrams, secondary structures have only non-crossing base pairs (arcs). Beyond RNA secondary structures are the RNA pseudoknot structures that allow for cross serial interactions [11]. Once such cross serial interactions are considered the question of a meaningful filtration arises, since the folding of unconstrained pseudoknot structures is NP-hard [12]. Based on several earlier studies of the genus of a pseudoknot single strand of RNA [13, 14, 15, 16], there are several meaningful filtrations of cross-serial interactions [17, 18, 19].

RNA interaction structures are diagrams over two backbones. Distinguishing internal and external arcs, the former being arcs within one backbone and the latter connecting the backbones, interaction structures can be represented by drawing the two backbones on top of each other, see Figure 2.

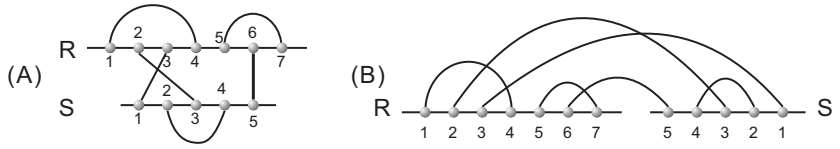

Figure 2: Diagram representation of an RNA-RNA interaction structure.

This paper will utilize a topological filtration to categorize RNA-complexes. While the basic concept of fat graphs employed here dates back to Cayley, the classification and expansion of pseudoknotted RNA structures in terms of topological genus of a fat graph or double line graph were first proposed by [17] and [20]. Fat graphs were applied to RNA secondary structures even earlier in [21] and [22]. The results of [17] are based on the matrix models and are conceptually independent. Genus, as well as other topological invariants of fat graphs were introduced and studied as descriptors of proteins in [23].

The approach undertaken here is combinatorial and follows [24]: starting with the diagram representation we inflate each edge, including backbone edges, into ribbons. As each ribbon has two sides and specifying a counter-clockwise rotation around each vertex, we obtain so called boundary cycles with a unique

orientation. It is clear that we have thus constructed a surface and its topological genus providing the filtration. Naturally there are many such ribbon graphs that produce the same topological surface (by gluing the two "complementary" sides of each ribbon), this is how we obtain the desired equivalence (complexity) classes of structures.

The idea of genus induction is an extension of the framework of [25, 26], who studied unicellular maps of genus g. In [27] a linear time algorithm for uniformly generating RNA structures of fixed topological genus was presented employing the results of [26]. In [28] this framework was extended to deal directly with RNA-shapes, i.e. enabling the uniform generation of finitely many shapes for fixed topological genus and to thereby extract key information from RNA databases.

In this contribution we derive the theory of RNA-RNA interaction structures by means of a new recursion. In the course of its construction we have to deal with the fact that it is not a "pure". This means it involves not only bicellular maps of lower genus but also disjoint pairs of unicellular maps. An additional novel feature is that our bijection is not *always* reducing topological genus. In essence we have the following alternative: we either reduce genus or we lose connectivity.

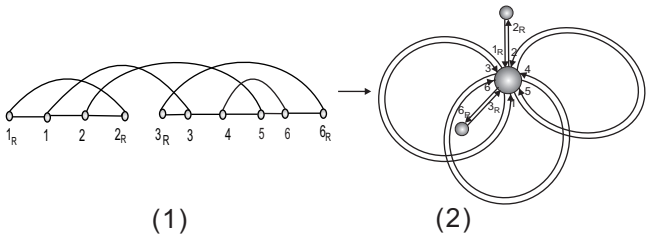

Figure 3: From a RNA-RNA interaction structures as diagrams to bicellular maps.

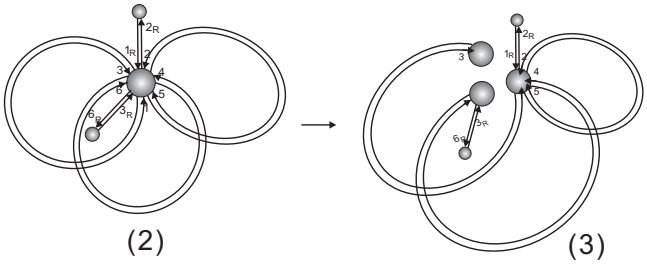

Figure 4: Slicing bicellular maps, see Section 3 for details. Slicing decreases genus by 1.

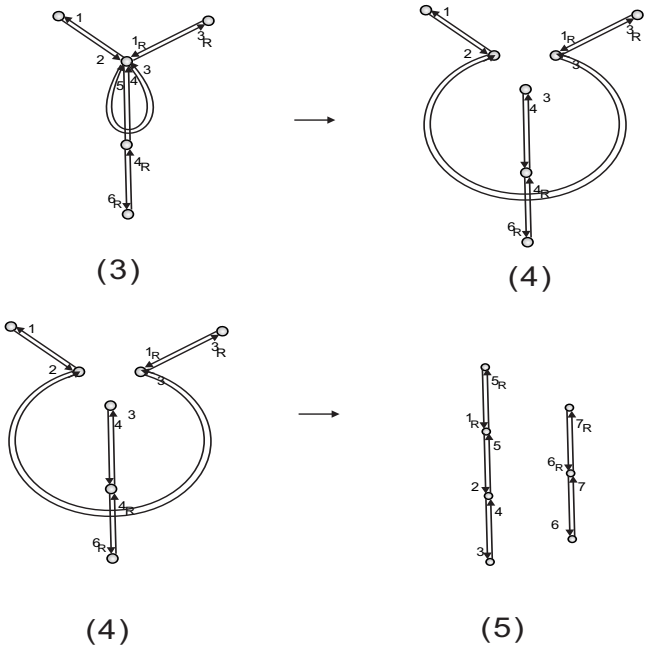

Figure 5: Further slicing into two plane trees.

The paper is organized as follows: In Section 2 we show that RNA-complexes are in one-to-one correspondence to such maps, namely those that are bicellular and planted, see Fig. 3. This correspondence allows us to perform all our constructions on maps and eventually recover the diagram thereafter. In Section 3

we study slicing and gluing of bicellular maps. We proceed by integrating the results of Section 3 into the main bijection and its combinatorial corollary in Section 4, see Fig. 4, 5. Finally we present the uniform generation algorithm in Section 5. Here the idea is to go back, i.e. we start from a pair of trees and successively rebuild the bicellular map. Finally, in Section 6, we discuss our results and show how to use our decomposition grammar to sample RNA-RNA interaction structures non-uniformly, employing a simplified loop-base model. This shows that the unambiguous grammar developed here has many applications and simply lifts the stochastic-context-free grammar approaches to secondary structures to structures with cross-serial interaction arcs. Various statistics about the loops and stacks in uniformly generated complexes of fixed topological genus are given and related to biological RNA-RNA interaction structures [29].

2. From RNA-complexes to bicellular maps and back

Definition 1. A diagram is a labeled graph over the vertex set $[n] = \{1, 2, \ldots, n\}$ represented by drawing the vertices $1, 2, \ldots, n$ on a horizontal line in the natural order and the arcs (i, j), where $i < j$, in the upper half-plane. The backbone of a diagram is the sequence of consecutive integers $(1, \ldots, n)$ together with the edges $\{\{i, i+1\} \mid 1 \leq i \leq n-1\}$. A diagram over b backbones is a diagram together with a partition of $[n]$ into b backbones, see Fig. 6.

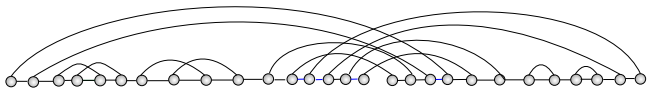

Figure 6: A 2-backbone diagram with 28 vertices and 14 arcs.

We shall distinguish backbone edges $\{i, i+1\}$ from arcs $(i, i+1)$, which we refer to as 1-arcs. Two arcs (i, j), (r, s), where $i < r$ are crossing if $i < r < j < s$ holds. Parallel arcs of the form $\{(i, j), (i+1, j-1), \cdots, (i+\ell-1, j-\ell+1)\}$

is called a stack, and ℓ is called the length of this stack. Furthermore, the particular arc, $(1, n)$, is called the rainbow.

Vertices and arcs of a diagram correspond to nucleotides and base pairs, respectively. For a diagram over b backbones, the leftmost vertex of each backbone denotes the $5'$ end of the RNA sequence, while the rightmost vertex denotes the $3'$ end. The particular case $b = 2$ is referred to as RNA interaction structures. Interaction structures are oftentimes represented alternatively by drawing the two backbones on top of each other.

We will add an additional "rainbow-arc" over each respective backbone and refer to these diagrams as *planted* 2-*backbone diagrams*, see Fig. 7.

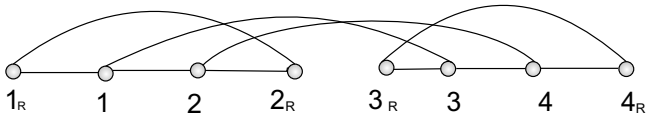

Figure 7: A planted 2-backbone diagram with its rainbow arcs $(1_R, 2_R), (3_R, 4_R)$.

The specific drawing of a diagram G in the plane determines a cyclic ordering on the half edges of the underlying graph incident on each vertex, thus defining a corresponding fat graph $\mathbb{G}$. The collection of cyclic orderings is called fattening, one such ordering on the half-edges incident on each vertex. Each fat graph $\mathbb{G}$ determines an oriented surface $F(\mathbb{G})$ which is connected if $\mathbb{G}$ is and has some associated genus $g(\mathbb{G}) \geq 0$ and number $r(\mathbb{G}) \geq 1$ of boundary components. Clearly, $F(\mathbb{G})$ contains G as a deformation retract. Without affecting topological type of the constructed surface, one may collapse each backbone to a single vertex with the induced fattening called the polygonal model of the RNA, see Fig 8.

We next prepare ourselves to study bicellular maps. To this end we discuss the idea behind general maps:

Definition 2. Let n be a positive integer. A map of size n is a triple $\mathfrak{m} = (\gamma, \alpha, \sigma)$ of permutations over $[1, 2n]$ such that:

• $\sigma\alpha = \gamma$,

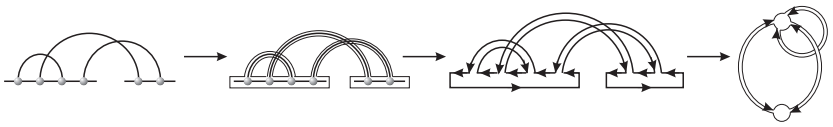

Figure 8: Inflation of a 2-backbone diagram and collapse of its 2 backbones to two vertices

• α is a fixed-point free involution (i.e. all its cycles have length 2).

As usual we write a permutation σ as a product of its cycles and denote the number of its cycles by $|\sigma|$. Suppose we are given a map $\mathfrak{m} = (\gamma, \alpha, \sigma)$, then the cycles of γ, α and σ are referred to as *faces*, *edges*, and *vertices*, respectively.

We can use fat graphs $\mathbb{G}$, which sometimes also called "ribbon graph", to give a graphical interpretation of maps. A fat graph is a multi-graph (with loops and multiple edges allowed), with a prescribed cyclic order (counter-clockwise) of the edges around each vertex.

Given a map $\mathfrak{m} = (\gamma, \alpha, \sigma)$, its associated fat graph $\mathbb{G}$ is the graph whose edges are given by the cycles of α, vertices by the cycles of σ, and the natural incidence relation $v \sim e$ if v and e share an element. Moreover, we draw each edge of $\mathbb{G}$ as a ribbon, where each side of the ribbon is called a half-edge; we decide which half-edge corresponds to which side of the ribbon by the convention that, if a half-edge h belongs to a cycle 4 of α and v of σ, then h is the right-hand side of the ribbon corresponding to e, when considered entering v. Furthermore, we draw the graph G in such a way that around each vertex v, the counter-clockwise ordering of the half-edges belonging to the cycle v is given by that cycle. Note that the cycles of of the permutation $\gamma = \sigma\alpha$ are interpreted as the sequence of half-edges visited when making a tour of the graph, keeping the graph on its left, see Fig. 9.

If the associated fat graph is connected, we call the map connected. A connected map $\mathfrak{m}$ can be embedded in a compact orientable surface, such that its complement is a disjoint union of simply connected domains (called the faces), and considered up to oriented homeomorphism. We can define the genus g of

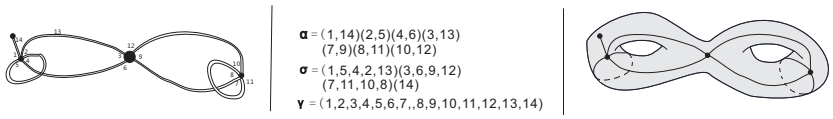

Figure 9: A map $\mathfrak{m} = (\gamma, \alpha, \sigma)$ left: fat graph representation, middle: permutation representation right: topological embedding.

the map $\mathfrak{m}$ by the genus of the surface. We can rewrite Euler's characteristic formula in terms of σ, γ and α as $|\sigma| + |\gamma| = |\alpha| + 2 - 2g$.

Now we are in position to discuss planted, bicellular maps.

Definition 3. A map $\mathfrak{b}_g = (\gamma, \alpha, \sigma)$ having n edges, genus g and boundary component $\gamma = \omega_1\omega_2 = (1, 2, 3, \cdots, k)(k+1, k+2, \cdots, 2n)$ is called bicellular if there exist some half-edge $x \in \omega_1$, such that $\alpha(x) \in \omega_2$.

Definition 4. A planted, bicellular map $\mathfrak{b}_g$ having n edges and genus g is a bicellular map $\mathfrak{b}_g = (\gamma, \alpha, \sigma)$, such that

$$\gamma = \omega_1\omega_2 = (1_R, 1, \ldots, m, m_R)((m+1)_R, (m+1)\ldots, 2n, (2n)_R),$$

where $|\gamma| = 2n + 4$ and α is a fixed-point free involution containing the cycles $(1_R, m_R)$ and $((m+1)_R, 2n_R)$. We refer to the latter as plants.

While bicellular maps are simply particular fat graphs, they naturally arise as the Poincaré dual of 2-backbone diagrams. That is, we have

Lemma 1. *There is a bijection between planted* 2*-backbone diagrams and planted bicellular maps.*

Proof. Given a planted 2-backbone diagram, we inflate the arcs, collapse each backbone into a single vertex, see Fig. 8. This produces a fat graph with two vertices (γ, α, σ). Next we consider the mapping (note $\gamma = \alpha \circ \sigma$):

$$\pi \colon (\gamma, \alpha, \sigma) \to (\sigma, \alpha, \alpha \circ \sigma).$$

π is evidently a bijection between fat graphs having two vertices and bicellular maps, see Fig. 10. The mapping is an instantiation of the Poincaré dual and

interchanges boundary components with vertices, preserving by construction topological genus. □

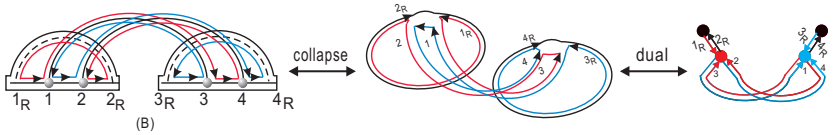

Figure 10: The Poincaré dual: from RNA complexes to bicellular maps and back.

3. Slicing and gluing in bicellular maps

Given a bicellular map $\mathfrak{b}_g$, the permutations σ and γ induce the following two linear orders $<_\gamma$ and $<_\sigma$ of half-edges: To define $<_\gamma$, we set $r_1 <_\gamma r_2$ for $r_1 \in \omega_1$ and $r_2 \in \omega_2$ and

$$r <_\gamma \omega_i(r) <_\gamma \cdots <_\gamma \omega_i^{k-1}(r), \qquad \text{for } \ i \in \{1,2\}.$$

Note that the minimal element here is the half-edge coming out from the first plant.

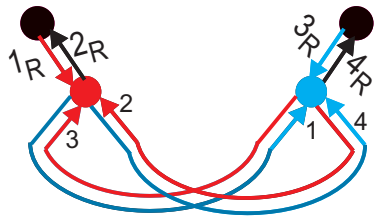

Figure 11: The two orders: $1_R <_\gamma 1 <_\gamma 2 <_\gamma 2_R <_\gamma 3_R <_\gamma 3 <_\gamma 4 <_\gamma 4_R$ and $1_r <_\sigma 3 <_\sigma 2$, $1 <_\sigma 4 <_\sigma 3_R$.

In order to define $<_\sigma$ we set for any vertex $v = \sigma_i$:

$$r <_\sigma \sigma_i(r) <_\sigma \cdots <_\sigma \sigma_i^k(r).$$

Let a_1 and a_2 be two distinct half-edges in $\mathfrak{b}_g$. Then $a_1 <_\gamma a_2$ expresses the fact that a_1 appears before a_2 in the boundary component ω_i or $a_1 \in \omega_1$ and

$a_2 \in \omega_2$. Suppose two half-edges a_1 and a_2 belong to the same vertex v. Note that v is a cycle which we assume to originate with the first half-edge along which one enters v travelling γ. Then $a_1 <_\sigma a_2$ expresses the fact that a_1 appears (counter-clockwise) before a_2.

Definition 5. A half-edge h is an *up-step* if $h <_\gamma \sigma(h)$, and a *down-step* if $\sigma(h) \leq_\gamma h$. h is called a *trisection* if h is a down-step and $\sigma(h)$ is not the minimum half-edge of its respective vertex.

This following lemma is the analogon to the trisection lemma of [26] for planted bicellular maps:

Lemma 2. *Any planted bicellular map, $\mathfrak{b}_g = (\gamma, \alpha, \sigma)$, has $2(g+1)$ trisections.*

Proof. Let n_+ and n_- denote the number of up-steps and down-steps in $\mathfrak{b}_g$. Then we have $n_+ + n_- = 2n + 4$, where n is the number of edges of $\mathfrak{b}_g$. Let i be a half-edge of $\mathfrak{b}_g$, and $j = \sigma^{-1}\alpha\sigma(i)$. Observe that we have $\sigma(j) = \gamma(i)$, and $\gamma(j) = \sigma(i)$. It is clear that if the tour of the map visits i before $\sigma(i)$, then it necessarily visits $\sigma(j)$ before j, see Fig. 12. We distinguish four cases:

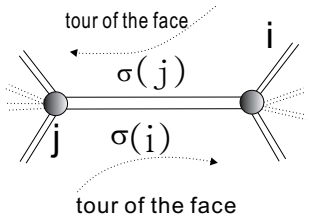

Figure 12: The main argument in the proof of the trisection lemma: the tour of the face visits i before $\sigma(i)$ if and only if it visits $\sigma(j)$ before j, unless $\sigma(i)$ or $\sigma(j)$ is the plants of the bicellular map.

First suppose $i <_\gamma \sigma(i) = \gamma(j)$, i.e. i is an up-step. Then $i <_\gamma \gamma(j)$ implies $\gamma(i) \leq_\gamma \gamma(j)$. Since $\gamma(i) = (\sigma\alpha)(i) = \sigma(j)$, this means $\sigma(j) \leq_\gamma \gamma(j)$. By definition of $<_\gamma$, this implies $\sigma(j) \leq_\gamma j$ since j is maximal with this property and $\sigma(j) \neq \gamma(j)$ (α has no fixed point). Accordingly, if i is an up-step, then j is a down-step.

Second, assume that $\sigma(i) = \gamma(j) \leq_\gamma i$ and that $\gamma(j)$ is not one of the two plants. In this case, $j <_\gamma \gamma(j)$ implies

$$j <_\gamma \gamma(j) = \sigma(i) \leq_\gamma i <_\gamma \gamma(i) = \sigma(j),$$

that is, $j <_\gamma \sigma(j)$, and j is consequently an up-step.

Third suppose that $\sigma(i) = \gamma(j) \leq_\gamma i$ and $\gamma(j) = 1_R$. Then $j = m_R = \alpha(1_R) = \alpha(\gamma(j)) = \sigma(j)$ and j is a down-step.

Fourth we suppose i is a down-step and $\gamma(j) = (m+1)_R$. Then $j = 2n_R$ is the biggest label of the half-edges. So j is always a down-step, i.e. $j \geq_\gamma \sigma(j)$.

Therefore we have proved that each edge, except of $(1_R, m_R)$ and $((m+1)_R, 2n_R)$ is associated to one up-step and one down-step. As a result there are exactly four more down-steps than up-steps, i.e. $n_- = n_+ + 4$, whence $n_- = n + 4$.

Since each vertex carries exactly one down-step which is not a trisection (its minimal half-edge), the total number of trisections equals $(n_- - v)$, where v is the number of vertices of $\mathfrak{b}_g$. Euler's characteristic formula, $v = n + 2 - 2g$, implies that the number of trisections is $n + 4 - (n + 2 - 2g) = 2g + 2$, whence the lemma. □

We next study the effect of slicing and gluing in bicellular maps.

Lemma 3. (slicing and gluing) *Suppose*

$$v := (a_1, h_2^1, \ldots, h_2^{m_2}, a_2, h_3^1, \ldots, h_3^{m_3}, a_3, h_1^1, \ldots, h_1^{m_1})$$

is a $\mathfrak{b}_g$-vertex and a_1, a_2, a_3 are intertwined, i.e. $a_1 <_\gamma a_3 <_\gamma a_2$ and $a_1 <_\sigma a_2 <_\sigma a_3$. Then slicing v via $\{a_1, a_2, a_3\}$ produces either a bicellular map $\mathfrak{b}_{g-1}$, or a pair of unicellular maps $(\mathfrak{u}_{g_1}, \mathfrak{u}_{g-g_1})$. Furthermore, slicing can be reversed by gluing via $\{a_1, a_2, a_3\}$.

Proof. Suppose

$$\gamma = \omega_1\omega_2 = ((1_{R_b}, 1_b, \ldots, m_b, m_{R_b})((m+1)_{R_b}, (m+1)_b, \ldots, (2n)_b, (2n)_{R_b})).$$

We distinguish the following two scenarios:

Case 1. a_1, a_2, a_3 are either contained in ω_1 or ω_2.

In this case it is clear that slicing preserves bicellularity. Indeed, suppose $\{a_1, a_2, a_3\} \in \omega_1$, then we can write the two faces as

$$(a_1, k_2^1, k_2^2, \cdots, k_2^{l_2}, a_3, k_1^1, k_1^2, \cdots, k_1^{l_1}, a_2, k_3^1, k_3^2, \cdots, k_3^{l_3})(k_4^1, k_4^2, \cdots, k_4^{l_4})$$

and slicing generates the boundary components

$$\begin{aligned}\hat{\gamma} = \hat{\omega}_1\hat{\omega}_2 =&(a_1, k_1^1, k_1^2, \cdots, k_1^{l_1}, a_2, k_2^1, k_2^2, \cdots, k_2^{l_2}, a_3, k_3^1, k_3^2, \cdots, k_3^{l_3})\\ &\cdot (k_4^1, k_4^2, \cdots, k_4^{l_4}).\end{aligned}$$

Hence $E_{con} := \{(x, \alpha(x)), | x \in \omega_1, \alpha(x) \in \omega_2\}$ is unaffected by slicing, which implies that the sliced map remains bicellular having two additional vertices. Since the number of edges remains constant Euler characteristic implies that the genus decreases by 1, see Fig. 13. *Case* 2. a_1, a_2, a_3 are not contained in

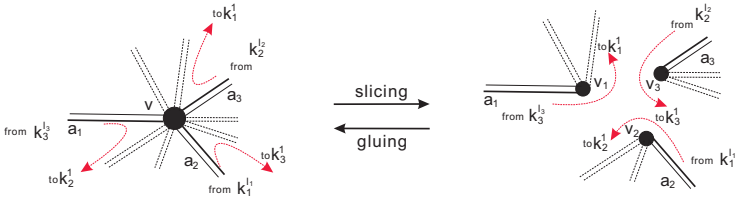

Figure 13: Gluing and slicing: case 1

either one of the boundary components. Clearly, $a_1 <_\gamma a_3 <_\gamma a_2$ implies that we have the alternative $\{a_1, a_3\} \in \omega_1, a_2 \in \omega_2$, or $\{a_1\} \in \omega_1, \{a_3, a_2\} \in \omega_2$.

In case of $\{a_1, a_3\} \in \omega_1, a_2 \in \omega_2$, we rewrite the two faces as

$$\gamma = \omega_1\omega_2 = (a_1, k_2^1, k_2^2, \cdots, k_2^{l_2}, a_3, k_1^1, k_1^2, \cdots, k_1^{l_1})(a_2, k_3^1, \cdots k_3^{l_3}).$$

We next consider the half-edges whose image is not the same for γ and $\hat{\gamma}$. These are $\{a_1, a_2, a_3\}$ and by construction we have

$$\hat{\gamma}(a_1) = \alpha\hat{\sigma}(a_1) = \alpha(h_1) = \alpha(\sigma(a_3)) = \gamma(a_3) = k_1^1.$$

Similarly, we have $\hat{\gamma}(a_2) = k_2^1$, and $\hat{\gamma}(a_3) = k_3^1$. Thus we arrive at

$$\hat{\gamma} = \hat{\omega}_1\hat{\omega}_2 = (a_1, k_1^1, k_1^2, \cdots, k_1^{l_1})(a_2, k_2^1, k_2^2, \cdots, k_2^{l_2}, a_3, k_3^1, k_3^2, \cdots, k_3^{l_3}).$$

Accordingly, slicing maps the set of half-edges $E_{mov} = \{k_2^1, k_2^2, \cdots, k_2^{l_2}, a_3\}$ into the second boundary component, ω_2. If $E_{con} \nsubseteq E_{mov}$, then we still have a bicellular map with genus $(g-1)$. However, if $E_{con} \subseteq E_{mov}$, then slicing produces a pair of unicellular maps $(\mathfrak{u}_1, \mathfrak{u}_2)$. Suppose that in this case $\mathfrak{u}_1$ has v_1 vertices, m_1 edges, and genus $2-2g_1 = v_1+1-m_1$. Then $\mathfrak{u}_2$ has $v_2 = v-v_1+2$ vertices, $m_2 = m-m_1$ edges and genus $2-2g_2 = v_2+1-m_2$, whence $g_2 = g-g_1$, see Fig. 14.

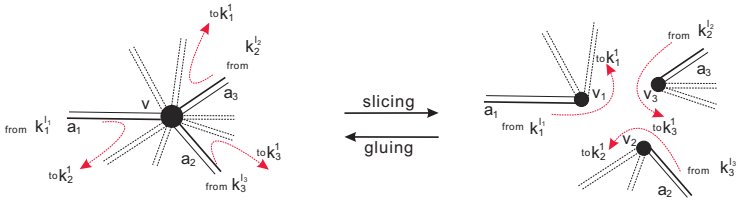

Figure 14: Gluing and slicing: case 2

The case $\{a_1\} \in \omega_1, \{a_3, a_2\} \in \omega_2$ is treated analogously. It is straightforward to verify that given $\{a_1, a_2, a_3\}$, slicing can be reversed by gluing. □

Definition 6. Given a bicellular map and a distinguished trisection τ at vertex v, we set a_1 to be the minimum half-edge of v, a_3 the half-edge following τ counter-clockwise and a_2 to be the smallest half-edge on the left of a_3, which is greater than a_3. Then slicing v via $\{a_1, a_2, a_3\}$, we have two scenarios: either a_3 is the minimum half-edge of its respective vertex, or not. In the former case τ is called type I and type II in the latter.

Let $D^I_{b,g+1}(n)$ denote the set of bicellular maps of genus $(g+1)$ with n edges and a distinguished trisection of type I. Let $B^3_g(n)$ denote the set of bicellular maps of genus g with n edges and three distinguished vertices. Finally, let $(U_{g_1}, U_{g+1-g_1})^3$ denote the set of pairs of unicellular maps whose sum of genera equals $(g+1)$, having combined n edges, such that not all distinguished vertices are contained exclusively in one map. We call such vertex configurations distributed.

Lemma 4. *There is a bijection*

$$\phi\colon D^{I}_{b,g+1}(n) \to B^{3}_{g}(n)\dot{\bigcup}(U_{g_1}, U_{g+1-g_1})^3.$$

Proof. Given a bicellular map $\mathfrak{b}_{g+1}(n)$ with vertex v, having a distinguished type I trisection, τ. Let ϕ be the slicing map of v via the half-edge set $C_\tau = \{a_1, a_2, a_3\}$, as in Def. 6. By construction, we have $a_1 <_\gamma a_3 <_\gamma a_2$, i.e. the half-edges are intertwined. From Lemma 3 we know that slicing produces either a bicellular map, $\mathfrak{b}_{g-1}(n)$, or alternatively a pair of unicellular maps $(\mathfrak{u}_{g_1}, \mathfrak{u}_{g+1-g_1})$. Furthermore slicing produces a triple (v_1, v_2, v_3), such that a_i is the minimum in the vertex v_i, respectively. In case of $(\mathfrak{u}_{g_1}, \mathfrak{u}_{g+1-g_1})$, such vertices are distributed. Accordingly, ϕ is well-defined.

We proceed by constructing the inverse of ϕ. To this end, let $\mathfrak{b}_g$ be a bicellular map of genus g with three distinguished vertices $\{v_1, v_2, v_3\}$, where

$$a_1 = \min_\gamma v_1 <_\gamma a_2 = \min_\gamma v_2 <_\gamma a_3 = \min_\gamma v_3.$$

Let $\chi(\mathfrak{b}_g)$ be the map obtained by the gluing of $\mathfrak{b}_g$ via $\{a_1, a_2, a_3\}$. By construction, a_1 remains to be the minimum half-edge of the vertex. $\sigma^{-1}(a_3)$ becomes a trisection which, by construction, is of type I and a_2 is by construction the smallest half-edge to the left of a_3 that is larger than a_3. Similarly, suppose we are given a pair $(\mathfrak{u}_{g_1}, \mathfrak{u}_{g+1-g_1})$ with three distinguished, distributed vertices v_1, v_2, v_3. By construction, gluing produces a bicellular map $\chi(\mathfrak{b}_g)$ with a distinguished trisection, $\sigma^{-1}(a_3)$. As slicing and gluing are inverse operations we have $\chi \circ \phi = \mathrm{id}$ and $\phi \circ \chi = \mathrm{id}$, whence ϕ is a bijection. □

Let $D^{II}_{b,g+1}(n)$ denote the set of bicellular maps of genus $(g+1)$ with n edges and a distinguished trisection of type II. Let $\nu_{\mathfrak{b},g}(n)$ be the set of 4-tuples $(\mathfrak{b}_g, v_1, v_2, \tau)$, where $\mathfrak{b}_g$ is a bicellular map of genus g with n edges and where v_1, v_2 and τ are two vertices and a trisection of $\mathfrak{b}_g$ such that:

$$\min_\gamma v_1 <_\gamma \min_\gamma v_2 <_\gamma \min_\gamma v(\tau).$$

Let $\kappa_{g+1}(n)$ be the set of 5-tuples $(\mathfrak{u}_{g_1}, \mathfrak{u}_{g+1-g_1}, v_1, v_2, \tau)$, where $\mathfrak{u}_{g_1}$ and $\mathfrak{u}_{g+1-g_1}$

are unicellular maps, with genus g_1 and $g+1-g_1$. Furthermore, v_1, v_2 and τ are distributed, i.e. not all three are contained in $\mathfrak{u}_{g_1}$ or $\mathfrak{u}_{g+1-g_1}$.

Then we have the following analogon of Lemma 4.

Lemma 5. *There exists a bijection* $\psi\colon D^{II}_{b,g+1}(n) \to \nu_{b,g}(n)\dot{\bigcup}\kappa_{g+1}(n)$.

Proof. Let $\mathfrak{b}_{g+1}(n)$ be a bicellular map of genus $(g+1)$ with vertex v and distinguished type II trisection τ. Let ψ be the slicing of v via $C = \{a_1, a_2, a_3\}$. Where the a_i are chosen as in Lemma 4. Lemma 3 guarantees that ψ generates either a bicellular map or a pair of unicellular maps and in the latter case v_1, v_2 and τ are distributed. However, slicing does not render a_3 as the minimum of v_3, since the trisection is type II. In fact, τ is again a trisection of v_3, since, by construction,

$$\hat{\sigma}(\tau) = a_3 \text{ and } a_3 <_{\hat{\gamma}} \tau$$

and there exist some $h_i \in \{h_3^1, h_3^2, \cdots h_3^{l_3}\}$ such that $h_i <_{\hat{\gamma}} a_3$. Therefore

$$\psi\colon D^{II}_{b,g+1}(n) \to \nu_{b,g}(n)\dot{\bigcup}\kappa_{g+1}(n)$$

is well-defined.

We proceed by specifying the inverse of ϕ, χ. Suppose we are given a bicellular map $\mathfrak{b}_g(n)$ or a pair of unicellular maps $(\mathfrak{u}_{g_1}, \mathfrak{u}_{g+1-g_1}) \in \kappa(n)$) with two vertices v_1, v_2 and a trisection τ. In case of $(\mathfrak{u}_{g_1}, \mathfrak{u}_{g+1-g_1})$ v_1, v_2 and τ are distributed. v_i and τ have the property

$$\min_{\hat{\gamma}} v_1 <_{\hat{\gamma}} \min_{\hat{\gamma}} v_2 <_{\hat{\gamma}} \min_{\hat{\gamma}} V(\tau).$$

Then we glue via the half-edges $a_1 = \min_{\hat{\gamma}} v_1, a_2 = \min_{\hat{\gamma}} v_2$ and $a_3 = \hat{\sigma}(\tau)$. By construction, a_3 is not minimal at v, whence τ is, after gluing, a type II trisection. Lemma 3 shows that the image of this gluing is contained in $D^{II}_{b,g+1}(n)$, whence

$$\chi\colon \nu_{\mathfrak{b},g}(n)\dot{\bigcup}\kappa_{g+1}(n) \to D^{II}_{b,g+1}(n)$$

is well-defined. By construction we have $\psi \circ \chi = \mathrm{id}$ and $\chi \circ \psi = \mathrm{id}$ and ψ is a bijection. □

4. The bijection

Let $D_{b,g}(n) = D^I_{b,g}(n) \cup D^{II}_{b,g}(n)$ be the set of bicellular maps of genus g with n edges and a distinguished trisection. Let $B_g(n)^t$ denote the set of a bicellular map of genus g with n edges and t distinguished vertices. Finally, let $(U_{g_1}(m), U_{g-g_1}(n-m))^t$, for $0 \leq g_1 \leq g$ denote the set of pairs of unicellular maps $(\mathfrak{u}_{g_1}(m), \mathfrak{u}_{g-g_1}(n-m))$ with m and $n-m$ edges and t distinguished, distributed vertices.

Theorem 1. *There exists a bijection*

$$\Xi_b \colon D_{b,g}(n) \to \left(\dot{\cup}_{p=0}^{g-1} B_p(n)^{2g-2p+1}\right) \dot{\bigcup} \left(\dot{\cup}_{g_1=0}^{p} \dot{\cup}_{m=0}^{n} \dot{\cup}_{p=0}^{g} \left(U_{g_1}(m), U_{p-g_1}(n-m)\right)^{2g-2p+3}\right). \tag{4.1}$$

Proof. Suppose we are given a bicellular map $\mathfrak{b} \in D_{b,g}(n)$ with distinguished trisection τ. Then we can recursively slice τ, as long as it remains a type II trisection. Clearly, τ must, after a finite number of slicings, become of type I and one more slicing resolves the latter into three distinguished vertices. In case of slicing into a pair of unicellular maps the distinguished vertices are distributed. Each slicing of a type II trisection produces vertices

$$\min_{\gamma} v_1 <_\gamma \min_{\gamma} v_2 <_\gamma \min_{\gamma} v_\tau,$$

and we write as $v_1 < v_2 < \tau$ for short. Since slicing does not affect the order of the half-edges between the plant and the minimum half-edge of the triple $\{a_1, a_2, a_3\}$, iterated slicings produces a sequence $v_1 < v_2 < \cdots < v_{2g-2p+1}$.

According to Lemma 4 and Lemma 5, the slicing of trisections of type II and I are indeed bijections. Furthermore, every slicing decreases topological genus by exactly 1 or 0(lose connectivity). As a result, after iteratively slicing the type II trisections, we obtain a type I trisection. Then one more slicing generates an element of

$$\left(\dot{\cup}_{p=0}^{g-1} B_p(n)^{2g-2p+1}\right) \dot{\bigcup} \left(\dot{\cup}_{g_1=0}^{p} \dot{\cup}_{m=0}^{n} \dot{\cup}_{p=0}^{g} \left(U_{g_1}(m), U_{p-g_1}(n-m)\right)^{2g-2p+3}\right)$$

and Ξ_b is accordingly well-defined. Clearly, Ξ_b is as the composition of bijections a bijection. $\square$

Let $\mathsf{U}_g(n)$ and $\mathsf{B}_g(n)$ denote the number of unicellular maps with genus g and n half-edges. Using the trisection lemma, Euler's formula, and Theorem 1, we obtain the following identity:

Corollary 1.

$$\begin{aligned} 2(g+1)\mathsf{B}_g(n) = &\sum_{i=0}^{g-1} \binom{n-2i}{2g-2i+1} \mathsf{B}_i(n) + \\ &\sum_{i=0}^{g} \sum_{g_1=0}^{i} \sum_{m=0}^{n} \sum_{k=1}^{2g-2i+2} \left(\binom{\mu_{g_1}}{k} \binom{\nu_{g_1,i}}{2g-2i+3-k} \mathsf{U}_{g_1}(m) \mathsf{U}_{i-g_1}(n-m) \right) \end{aligned} \tag{4.2}$$

where

$$\begin{aligned} \mu_{g_1} &= m+1-2g_1 \\ \nu_{g_1,i} &= n-m-2i+2g_1+1. \end{aligned}$$

Proof. Consider $\mathfrak{b}_i$ with n arcs. According to Euler's formula, we have $v = n-2i$ vertices. Since every gluing increases genus by 1, there are exactly $(g-i)$ gluings to derive genus g. The first gluing requires 3 vertices and generates a type I trisection. Every following step requires 2 vertices, whence we choose $2(g-i)+1$ vertices. This interprets the binomial coefficients $\binom{n-2i}{2g-2i+1}$. Similarly, if we are given a pair $(\mathfrak{u}_{g_1}, \mathfrak{u}_{i-g_1})$, Suppose $\mathfrak{u}_{g_1}$ has m edges and the other map has $(n-m)$ arcs. Then $\mathfrak{u}_{g_1}$ and $\mathfrak{u}_{i-g_1}$ have $\mu_{g_1} = m-2g_1+1$, $\nu_{g_1,i} = n-m-2(i-g_1)+1$ vertices, respectively. Every gluing step increases genus by 1 except one step which connects the two unicellular maps to a bicellular map, preserving the genus. Thus $(g-i+1)$ gluings generate genus g and we need to choose $2(g-i+1)+1$ distributed vertices. Suppose we choose k vertices on $\mathfrak{u}_{g_1}$, then k must satisfy $1 \leq k \leq 2(g-i+1)$ and the other vertices are selected from $\mathfrak{u}_{i-g_1}$. This interprets the binomial coefficients $\binom{\mu_{g_1}}{k}\binom{\nu_{g_1,i}}{2g-2i+3-k}$, whence the corollary. $\square$

Any bicellular map $\mathfrak{b}_g$ together with one of its $2(g+1)$ trisections is mapped via Ξ_b into a bicellular map of lower genus or a pair of unicellular maps. Note

that either the topological genus decreases by at least one or we lose connectivity and decompose into a pair of unicellular maps. From [26], we know that a unicellular map can be iteratively sliced into a planar tree. Therefore, we have

Corollary 2. *Any bicellular map can be sliced into a pair of planar trees and we have*

$$\begin{aligned}
\mathsf{B}_g(n) = & \sum_{g_0<g_1<g_i<\cdots<g_{r-1}<g_r=g} \sum_{b=1}^{r} \sum_{m=0}^{n} \\
& \left(\sum_{l\geq 0}^{b-1} \binom{b-1}{l} \prod_{i=i_v\in I_l} \frac{1}{2g_{i,A}} \binom{m+1-2g_{(i_{v-1},A)}}{2(g_i-g_{i-1})+1} \times \right. \\
& \prod_{i=j_v\in J_{b-l-1}} \frac{1}{2g_{i,B}} \binom{n-m+1-2g_{(j_{v-1},B)}}{2(g_i-g_{i-1})+1} \times \\
& \left. \left(\sum_{k\geq 1}^{2(g_b-g_{b-1})+1} \frac{1}{2g_b+2} \binom{m+1-2(g_{b-1,A})}{k} \binom{n-m+1-2(g_{b-1,B})}{2(g_b-g_{b-1})+1-k} \right) \right) \\
& \times \prod_{i=b+1}^{r} \frac{1}{2g_i+2} \binom{n-2g_{i-1}}{2(g_i-g_{i-1})+1} \epsilon_0(m)\epsilon_0(n-m),
\end{aligned} \tag{4.3}$$

where

$$\begin{aligned}
I_l &= \{i_1, i_2, \cdots i_l\}, \quad J_{b-l-1} = \{j_1, j_2, \cdots j_{b-l-1}\}, \\
g_{b-1,A} &= \sum_{i_x=i_1}^{i_l} (g_{i_x} - g_{i_x-1}), \quad g_{b-1,B} = g_{b-1} - g_{b-1,A}, \\
g_{(i_{v-1},A)} &= \sum_{i_x=i_1}^{i_{(v-1)}} (g_{i_x} - g_{i_x-1}),\ g_{(j_{v-1},B)} = \sum_{j_y=j_1}^{j_{(v-1)}} (g_{j_y} - g_{j_y-1}),
\end{aligned}$$

and $\epsilon_0(n)$ denotes the number of plane trees with n edges.

5. Uniform generation

In this section, we show how to generate a bicellular map of given genus g over n edges with uniform probability.

Here is the key idea: according to Corollary 2, any bicellular map decomposes into to a pair of plane trees $(\mathfrak{u}_0(m), \mathfrak{u}_0(n-m))$. Recruiting the inverse to

slicing, the gluing, we can recover $\mathfrak{b}_g$. As each bicellular map is generated with multiplicity $2(g+1)$, see Corollary 1, we can employ our bijection to uniformly generate bicellular maps of fixed topological genus g.

We first give the definition of glue path.

To this end, let $\mathfrak{b}_g$ denote a bicellular map of genus g having n edges, let $\mathfrak{p}_g$ denote a pair of unicellular maps of genus sum g and let $\mathfrak{m}$ denote a map.

Definition 7. A glue path starting from $\mathfrak{p}_0$ to $\mathfrak{b}_g$, is a sequence

$$\left((\mathfrak{m}^0 = \mathfrak{p}_{g_0=0}, j_0 = 0), \ldots, (\mathfrak{m}^i, j_i), \ldots, (\mathfrak{m}^b = \mathfrak{b}_{g_b}, j_b = 1), \ldots, (\mathfrak{m}^r = \mathfrak{b}_g, j_r = 1)\right),$$

where $j_i \in \{0, 1\}$ is a flag, an indicator variable for connectivity. b is the first step where j_i switches to 1. The corresponding sequence

$$\left((g_0 = 0, j_0 = 0), \ldots, (g_i, j_i), \ldots, (g_r, j_r = 1)\right).$$

is called the signature of the glue path.

We shall generate $\mathfrak{b}_g(n)$ in two steps: first we construct a pair of planar trees $\mathfrak{p}_0 = (\mathfrak{u}_0(m), \mathfrak{u}_0(n-m))$ with n edges with uniform probability. There are $\sum_{m=0}^{n} \epsilon_0(m)\epsilon_0(n-m)$ such pairs. Second, starting from this pair, we generate a glue path to the target genus.

It is well-known how to generate a plane tree with n edges in linear time [30]. For every pair $\mathfrak{p}_0 = (\mathfrak{u}_0(m), \mathfrak{u}_0(n-m))$, we next generate a glue path with uniform probability as follows.

For a given pair of unicellular maps $\mathfrak{p}_0$ and target genus g, we first construct all signatures.

For every such path we have $(g_0, j_0) = (0, 0)$ and $(g_r, j_r) = (g, 1)$. We can construct the signatures inductively. The induction basis is trivial, as for the step, suppose we have arrived at (g_i, j_i), where $0 \leq g_i \leq g$ and $j_i = 0$ or 1. If $j_i = 0$, then we can generate either $\{(g_i + 1, 0), (g_i + 2, 0), \ldots (g, 0)\}$ or $\{(g_i, 1), (g_i + 2, 1), \ldots (g, 1)\}$. If $j_i = 1$, then we obtain $\{(g_i + 1, 1), (g_i + 2, 1), \ldots (g, 1)\}$. Since the initial and final tuples are fixed, after finitely many iterations we can thereby generate all the signatures, see Fig. 15.

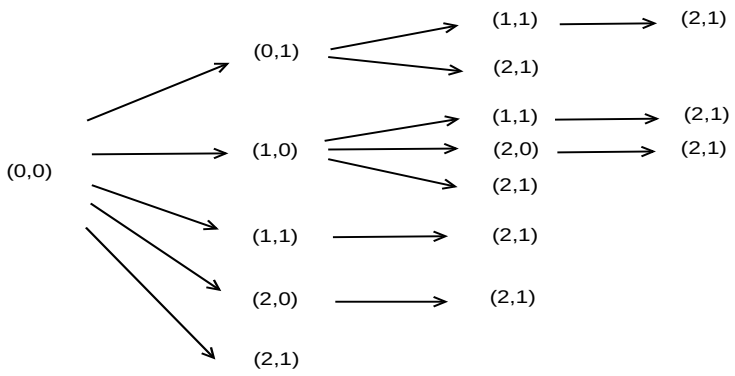

Figure 15: the signatures for target genus 2.

Every signature path has a probability. It is given by the number of glue paths from $(\mathfrak{p}_0, 0)$ to $(\mathfrak{m}_{g_{i+1}}, j_{i+1})$ having this signature, normalized by the total number of glue paths.

We arrive at

$$\begin{aligned}&\mathbb{P}((g_{i+1}=t, j_{i+1}) \mid (g_0,0)\ldots,(g_i,j_i),(g,1)) \\ &= \frac{\sum_{(t_0=g_0,0)\ldots,(t_i=g_i,j_i),(t_{i+1}=g_{i+1},j_{i+1})\cdots(t_r=g,1)} \Omega_{i+1}}{\sum_{(t_0=g_0,0)\ldots,(t_i=g_i,j_i)\cdots(t_r=g,1)} \Omega_i},\end{aligned} \tag{5.1}$$

where the Ω_i denotes the sum of weights over all the signatures that contain (g_i, j_i). We can thus compute all transition probabilities of sates of signatures state and derive the transition matrix M.

Given M, we can construct a glue path as follows:

Suppose we are at step i and we have constructed a map $\mathfrak{m}^i$ with (g_i, j_i). Then (g_{i+1}, j_{i+1}) can be derived via M, by the process `NextTuple`.

Next we select the vertices and gluing accordingly. Let (A, B) be the pair of plane trees. Suppose we have the tuple (m_{g_i}, j_i), `NextTuple` produces (g_{i+1}, j_{i+1}) and if $j_{i+1} = 0$, then we select the vertices "locally" on one of the unicellular maps via

$$\mathbb{P}(V_{i,1}) = \frac{1}{\binom{m+1-2g_{i,A}}{2(g_{i+1}-g_i)+1}}$$

or

$$\mathbb{P}(V_{i,2}) = \frac{1}{\binom{n-m+1-2g_{i,B}}{2(g_{i+1}-g_i)+1}}.$$

In case of $j_{i+1} = 1$ and $j_i = 0$, we need to choose the vertices distributed i.e.

$$\mathbb{P}(V_{i,3}) = \frac{1}{\sum_{k=1}^{2(g_{i+1}-g_i)} \binom{m+1-2g_{i,A}}{k} \binom{n-m+1-2(g_i-g_{i,A})}{2(g_{i+1}-g_i)+1-k}}.$$

Finally in case of $j_{i+1} = 1$ and $j_i = 1$ we can choose vertices arbitrarily, i.e.

$$\mathbb{P}(V_{i,4}) = \frac{1}{\binom{n-2g_i}{2(g_{i+1}-g_i)+1}}.$$

We refer to the above three cases of local, distributed and free vertex selection as `SelectVertex1`, `SelectVertex2` and `SelectVertex3`.

After the sequence of vertices V_i is selected, a bicellular map $\mathfrak{b}^{i+1}$ is constructed by the process `Glue`. Notice that in accordance with Theorem 1 after every application of `Glue`, we normalize by $2g_i$, for $j_i = 0$, or $2g_i + 2$ in case of $j_i = 1$. We present the pseudocode of the procedures in Algorithm 1.

Since the target genus is fixed constant and since the intermediate genera are monotone, the while-loop of Algorithm 1 is executed only a constant number of times. Using appropriate memorization techniques, `NextTuple` and `Glue` can be implemented in constant time. Furthermore `SelectVertex1`, `SelectVertex2` and `SelectVertex3` have linear run-time complexity. Thus, combined with a linear time sampler for planar trees, our approach allows for the uniform generation of random 2-backbone matchings in $O(n)$ time.

We accordingly obtain

Corollary 3. *Algorithm 1 generates uniformly bicellular maps.*

Proof. First, we generate the plane trees uniformly. Second, by construction, every transition from $\mathfrak{m}^i$ to $\mathfrak{m}^{i+1}$ is a bijection by Theorem 1 and uniform after normalizing by $2(g_{i+1} + 1)$, whence the corollary. □

Now we can extend our result in order to generate 2-backbone diagrams of genus g with uniform probability. The idea is as follows: first we uniformly

Algorithm 1 Generation of glue path for bicellular maps

1: `UniformBi-Matching` $(\mathfrak{m}^0 = \mathfrak{p}^0, Targettuple)$
2: $i \leftarrow 0, j \leftarrow 0$
3: **while** $(g_i, j_i) \neq Targettuple$ **do**
4: $(g_{i+1}, j_{i+1}) \leftarrow$ `NextTuple` $((g_0, 0) \ldots, (g_i, j_i), Targettuple)$
5: **if** $j_{i+1} = 0$ **then**
6: $V_i \leftarrow$ `SelectVertex1` $(\mathfrak{m}^i, 2(g_{i+1} - g_i) + 1)$
7: **else if** $j_{i+1} = 1 \& j_i = 0$ **then**
8: $V_i \leftarrow$ `SelectVertex2` $(\mathfrak{m}^i, 2(g_{i+1} - g_i) + 1)$
9: **else if** $j_{i+1} = 1 \& j_i = 1$ **then**
10: $V_i \leftarrow$ `SelectVertex3` $(\mathfrak{m}^i, 2(g_{i+1} - g_i) + 1)$
11: **end if**
12: $\mathfrak{m}^{i+1} \leftarrow$ `Glue` $(\mathfrak{m}^i, V_i)$, $i \leftarrow i + 1$
13: **end while**
14: **return** $\mathfrak{m}^i$

generate a 2-backbone matching of genus g with n arcs. Then we choose $(\ell - 2n)$ unpaired vertices and insert them into the matching.

Let $\mathbb{P}_d(t = n|\ell, g)$ denote the probability of the 2-backbone diagram of length ℓ and genus g having exactly n arcs, $0 \leq n \leq \lfloor \ell/2 \rfloor$. Let $\delta_g(\ell)$ denote the number of 2-backbone diagrams of genus g over ℓ vertices. Furthermore, let $\delta_g(\ell, n)$ denote the number of diagrams of genus g over ℓ vertices having exactly n arcs, $2n \leq \ell$. Denote the number of 2-backbone matchings with n arcs by $\xi_g(n)$. Then

$$\begin{aligned} \delta_g(\ell, n) &= \binom{\ell}{\ell - 2n} \xi_g(n) \\ \delta_g(\ell) &= \sum_{n=0}^{\lfloor l/2 \rfloor} \delta_g(\ell, n) = \sum_{n=0}^{\lfloor l/2 \rfloor} \binom{\ell}{\ell - 2n} \xi_g(n) \end{aligned}$$

and $\mathbb{P}_d(t = n|\ell, g) = \delta_g(\ell, n)/\delta_g(\ell)$.

This leads to Algorithm 2, which generates uniformly diagrams of length ℓ of genus g. Accordingly, a 2-backbone diagram of genus g over ℓ vertices with

exactly n arcs is generated, which we denote by $D_g^2(l)$.

Algorithm 2

1: `Uniform2BackboneDiagram` $(\ell, TargetGenus)$

2: $n \leftarrow$ `NumberofArcs`(ℓ, g)

3: $\mathfrak{p}_0 \leftarrow$ `UnifomTrees`(n)

4: $\mathfrak{b}_g \leftarrow$ `UniformBi` $-$ `Matching`$(\mathfrak{p}_0, TargetGenus)$

5: $D_g^2(l) \leftarrow$ `InsertUnpairedVertices`$(\mathfrak{b}_g, \ell)$

6: **return** $D_g^2(l)$

As for the subroutines:

- `NumberofArcs` returns n with probability $\mathbb{P}_d(t = n|\ell, g)$ and accordingly gives the number of arcs in 2-backbone diagram of length ℓ,
- `UniformTrees` uniformly generates a pair of planer trees with a total of n arcs,
- `UniformBi-Matching` generates a 2-backbone matching of genus g with n arcs,
- `InsertUnpairedVertices` selects $(\ell - 2n)$ vertices from ℓ vertices as to be unpaired and inserts them.

The result of some experiments conducted in connection with the generation of random matchings and diagrams are displayed in Fig. 16.

6. Discussion

We derived a uniform generation algorithm for RNA-RNA interaction structures of fixed topological genus. The algorithm is very fast having only linear time complexity. It allows immediately to obtain an abundance of statistical data on these structures.

In the following we shall consider biased sampling of RNA-RNA interaction structures. The bias is obtained by employing a simplified version of extending

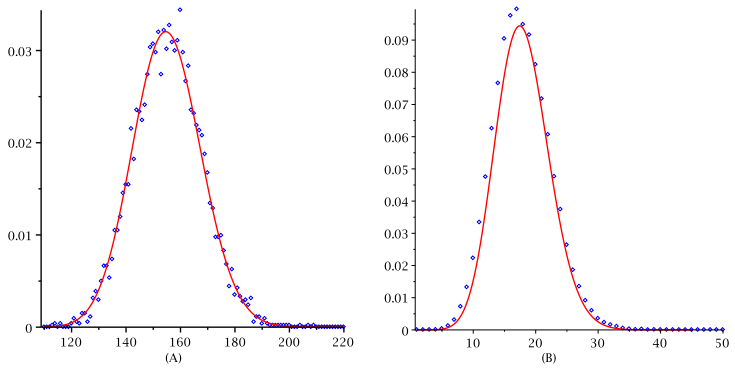

Figure 16: Uniform generation: (A) matchings over 2 backbones, $n = 10$, $g = 1$. We generate $N = 10^6$ matchings and display the frequencies of their multiplicities (blue dots) together with the Binomial coefficient of the uniform sampling. (B) The analog of (A) for diagrams, i.e. incorporating isolated vertices. Here we have $n = 10$, $g = 1$, We generate $N = 10^6$ diagrams and display the frequencies of their multiplicities (blue dots) together with the Binomial coefficient of the uniform sampling.

the bio-physical loop-energy model of RNA secondary structures to RNA-RNA interaction structures. Here we restrict ourselves to the case of genus 0 structures, but the treatment of higher genera structures is straightforward from here. Note that genus 0 interaction structures exhibit in general cross serial interactions between their two backbone, i.e. exhibit crossing arcs.

RNA structures are, due to the biophysical context, subject to specific constraints with respect to their free energy [31]. The latter is oftentimes modelled as a function of the loops of the underlying RNA structure [31]. This goes back to Waterman *et al.* [32, 33, 34, 35, 36] who realized that the classic secondary structure recursion is compatible with the loop energy model. It is interesting to note that these loops actually correspond to faces in the fat graph model, that is boundary components. This phenomenon naturally extends to structures over any number of backbones and any topological genus. Their loops are also just topological boundary components and the framework extend in a natural way, see Fig. 21. In case of RNA secondary structures, we find essentially three types

of loops: hairpin loops, interior loops (including helices and bulge loops) and multi-loops. The Poincaré duality described in Lemma 1 interchanges boundary components and vertices, whence we have the following correspondences

- hairpin loops and vertices of degree one,
- interior loops and vertices of degree two and
- multi-loops and vertices of degree greater than two, see Fig. 17.

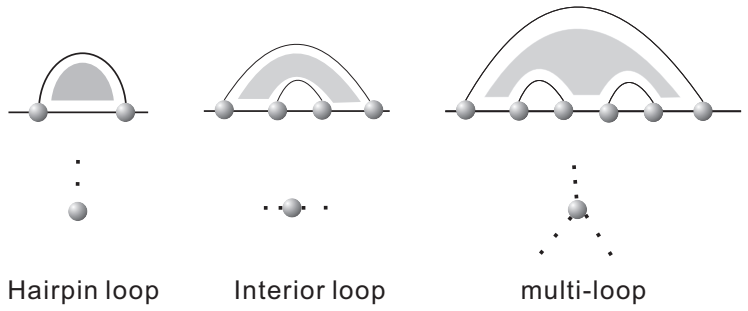

Figure 17: The three loop types in the RNA secondary structures.

Let B_g denote an RNA-RNA interaction structure having length l, n arcs and genus g. Lemma 1 associates to this diagram and a bicellular map, denoted by $\mathfrak{b}_g$.

In Section 4 we discussed that given a bicellular map $\mathfrak{b}_g$ together with a distinguished trisection, a finite number of vertex-slicings produces a pair of plane trees together with a collection of labeled vertices.

We showed in Theorem 1 that any such slicing is reversible, whence the decomposition via vertex slicings is unique. This means that we actually have derived an unambiguous grammar that decomposes any RNA-RNA interaction structure of fixed topological genus into an (ordered) pair of secondary structures with some labeled loops.

Then the energy of such a structure$\eta(B_g)$ is given by

$$\eta(B_g) = n \cdot b + \eta(S_1^p) + \eta(S_2^q) + L_g,$$

where b is the energy contribution of an arc. S_1^p (S_2^q) is a secondary structure with p (q) marked loops. Furthermore, $\eta(S_1^p) = \sum_{X_1} L^{X_1}$ and $\eta(S_2^q) = \sum_{X_2} L^{X_2}$, where X_1 (X_2) are the set of unmarked loops in S_1 (S_2). L_g represents the energy contribution of the labeled loops and the contribution of the gluing path.

To illustrate what happens here, let us have a look at the case of $g = 0$. Suppose we are given a genus 0 matching over 2 backbones, B_0. According to Theorem 1 its dual bicellular map, $\mathfrak{b}_0$, corresponds to a pair of trees and together with three labeled vertices, denoted by (T_1, T_2). The corresponding two pseudoknot-free secondary structure with three labeled boundary components are denoted by $(S_1 \cup S_2)$. Thus we have $\eta(B_0) = \eta(S_1^1) + \eta(S_2^2) + L_0 = \eta(S_1^1) + \eta(S_2^2) + L^{mul} + \epsilon$, where $L_0 = L^{mul} + \epsilon$, since the three vertices after gluing will form a vertex of degree at least 3, corresponding to a multi-loop. Finally, ϵ can be regarded as the contribution of the particular type of pseudoknot being glued. The situation is particularly transparent for $g = 0$, since there are only two shapes E and F, where a shape is a diagram without unpaired vertices and 1-arcs in which all stacks (parallel arcs) have size one. These two shapes are depicted in Fig. 18 and Fig. 19, where we show in addition these two shapes and the pair of secondary structures with three labeled boundary components they slice into. we accordingly derive

$$\eta(E) = \eta(S_1^1)+\eta(S_2^2)+L_0 = L^{mul}+\epsilon, \quad \eta(F) = \eta(S_1^1)+\eta(S_2^2)+L_0 = 2L^{mul}+\epsilon.$$

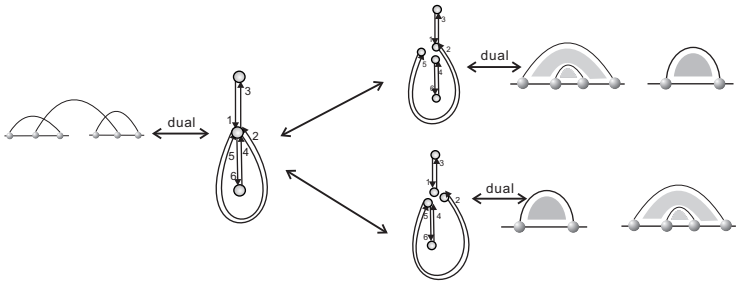

Figure 18: the E-shape and the pair of secondary structures it slices into.

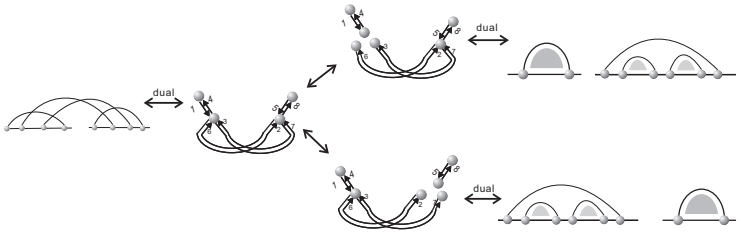

Figure 19: the F-shape and the pair of secondary structures it slices into.

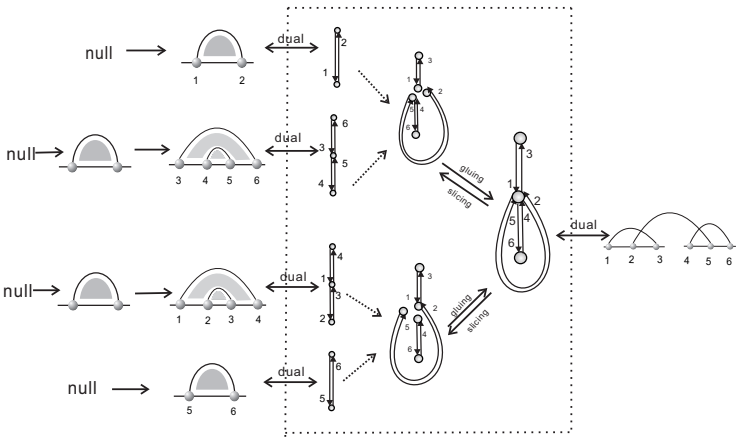

Figure 20: Howto generate the E-shape. left: generating the two secondary structures, middle: gluing the corresponding dual maps of the latter two, right: E-shape obtained by dualizing again.

It is important to note that our sampler is based on a two literally "orthogonal" compositions. The first is inductive in length and adds either unpaired vertices or arcs. There is no topological "complexity" in the structures. Point in case: any RNA secondary structures has genus 0. The second is inductive in either topological genus or connectedness but adds neither vertices nor arcs. This induction is novel and substantially different from length based induction, See Fig. 20 for an illustration for the generation of E-shape of genus 0.

We shall proceed and study several statistics of loops in RNA-RNA interaction structures, see Fig. 21.

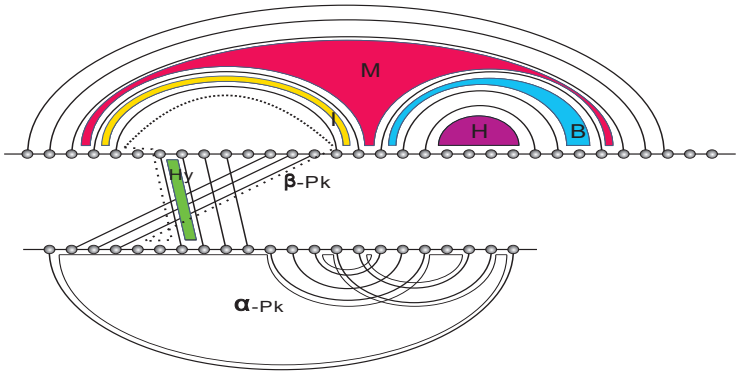

Figure 21: RNA-RNA interaction structure and its loops: multi-loop, M, interior loop, I, hairpin loop, H, bulge loop, B, exterior stack H_y, pseudoknot loop over one backbone, α-Pk, and pseudoknot loop over two backbones β-Pk.

We first present the distribution of loop types in interaction structures of genus g, see Fig. 21. We shall distinguish loops that contain only edges with endpoints on one backbone (α-loops) and those that contain also edges connecting the two backbones (β loops), see Fig. 22 and 23.

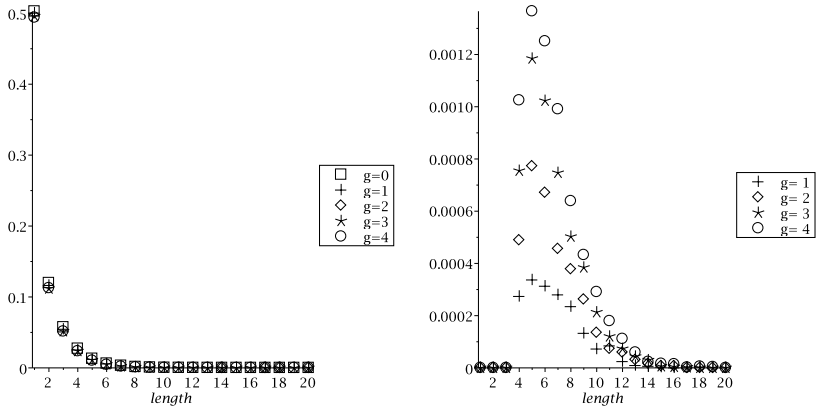

Figure 22: The length of α-loops in uniformly generated, genus filtered, RNA structures. Data are obtained from 10^5 interaction structures of length 500 with genus ranging from 0 to 4, respectively. left: distribution of standard α-loops, where the x-axis represents the length of the boundary component and y-axis denotes frequency. right: distribution of pseudoknot α-loops.

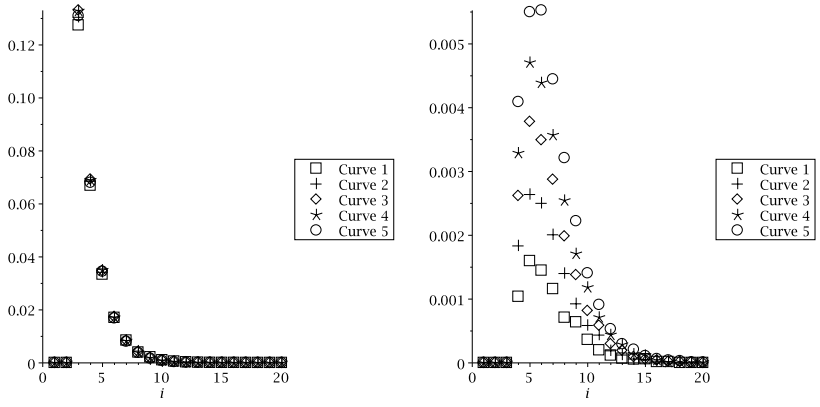

Figure 23: The length-distribution of β-loops, sample and set-up as in Fig. 22.

Next we depict the distribution stack-length of uniform versus biological RNA-RNA interaction structures obtained from [29]. In both distributions we observe that lower stack length appears with high probability, see Fig. 24.

Finally we present the distribution of β stacks versus that of both, α- and β-stacks, see Fig. 25.

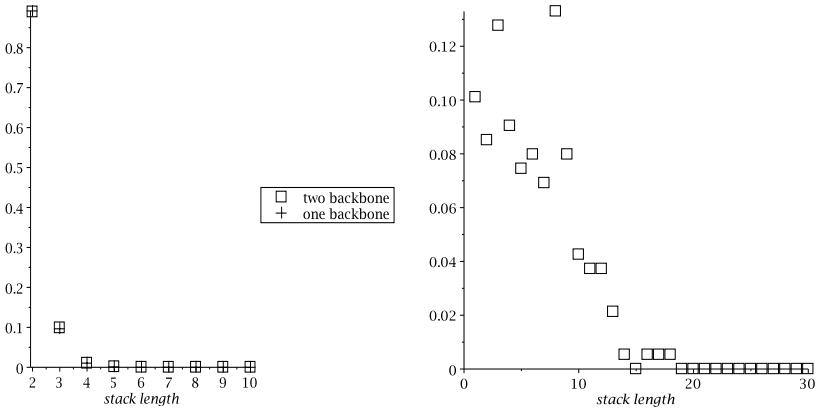

Figure 24: left: The distribution of the stack-length of 5×10^4 uniformly generated genus 1 interaction structures of length 500. right: Distribution of the stack length of biological structures [29].

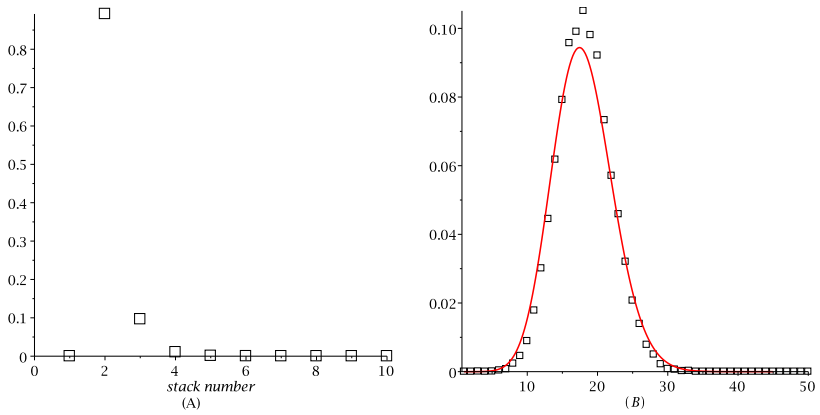

Figure 25: β-stacks versus all stacks: sampling 10^5 interaction structures of genus 1, length 500, we display: left: the distribution of the number of β stacks, right: the distribution of the number of all stacks.

We have in Theorem 1 the blueprint for a novel, multiple-context free gram-

mar, generating unambiguously RNA-RNA interaction structures of genus g. This grammar is genuinely topological and can be used for a variety of applications. For instance it can be tailored to produce not uniform but biological interaction structures by means of a training set taken from a database of RNA-RNA interaction structures. As is standard in stochastic-(multiple) context free grammars, this training set provides the probabilities of the rules. It would be then possible to statistically validate the finding by comparing the derived loop-size statistics from biased sampling with that of biological interaction structures. Another interesting application would arise in the context of functional anotation, where via sequencing sites that encode specific pseudoknot RNA like telomerases. The key objective is the development of local descriptors, as suitable input for efficient, genome-wide search, which requires deeper, conceptional understanding of RNA pseudoknots.

Author contributions Hillary S.W. Han obtained an arithmetic proof of Eq. (4.2) based on [26] and [37] and generated Figures (3(2), 4, 5(4), 8, 9, 10, 13, 14, 21). Eq. (4.3) was jointly derived by all authors. Benjamin M.M.Fu and Christian M. Reidys derived the bijections, designed the algorithms, the statistical results and wrote the paper.

Acknowledgements We wish to thank Fenix W.D. Huang and Thomas J.X. Li for discussions. This work is funded by the Future and Emerging Technologies (FET) programme of the European Commission within the Seventh Framework Programme (FP7), under the FET-Proactive grant agreement TOP-DRIM, FP7-ICT-318121.

References

[1] F. Narberhaus, J. Vogel, Sensory and regulatory RNAs in prokaryotes: A new german research focus, RNA biology 4 (3) (2007) 160–164.

[2] M. T. McManus, P. A. Sharp, Gene silencing in mammals by small interfering RNAs, Nature reviews genetics 3 (10) (2002) 737–747.

[3] D. Banerjee, F. Slack, Control of developmental timing by small temporal RNAs: a paradigm for RNA–mediated regulation of gene expression, Bioessays 24 (2) (2002) 119–129.

[4] J.-P. Bachellerie, J. Cavaillé, A. Hüttenhofer, The expanding snoRNA world, Biochimie 84 (8) (2002) 775–790.

[5] R. Benne, RNA–editing in trypanosome mitochondria, Biochimica et Biophysica Acta (BBA)-Gene Structure and Expression 1007 (2) (1989) 131–139.

[6] J. F. Kugel, J. A. Goodrich, An RNA transcriptional regulator templates its own regulatory RNA, Nature chemical biology 3 (2) (2007) 89–90.

[7] D. Kleitman, Proportions of Irreducible Diagrams, Studies in Appl. Math. 49 (1970) 297–299.

[8] R. Nussinov, G. Pieczenik, J. R. Griggs, D. J. Kleitman, Algorithms for loop matchings, SIAM Journal on Applied mathematics 35 (1) (1978) 68–82.

[9] M. S. Waterman, Secondary structure of single–stranded nucleic acids, Adv. Math. (Suppl. Studies) 1 (1978) 167–212.

[10] M. S. Waterman, Combinatorics of RNA hairpins and cloverleaves, Studies Appl. Math 60 (1978) 91–96.

[11] E. Rivas, S. R. Eddy, A dynamic programming algorithm for RNA structure prediction including pseudoknots, J. Mol. Biol. 285 (1999) 2053–2068.

[12] R. B. Lyngsø, C. N. Pedersen, Pseudoknots in RNA secondary structures, in: Proceedings of the fourth annual international conference on Computational molecular biology, ACM, 2000, pp. 201–209.

[13] G. Vernizzi, H. Orland, Large–N random matrices for RNA folding, Acta PhysicA PolonicA Series B 36 (9) (2005) 2821.

[14] G. Vernizzi, H. Orland, A. Zee, Enumeration of RNA structures by matrix models, Physical review letters 94 (16) (2005) 168103.

[15] M. Bon, G. Vernizzi, H. Orland, A. Zee, Topological classification of RNA structures, Journal of molecular biology 379 (4) (2008) 900–911.

[16] J. E. Andersen, R. C. Penner, C. M. Reidys, M. S. Waterman, Topological classification and enumeration of RNA structures by genus, Journal of mathematical biology (2011) 1–18.

[17] H. Orland, A. Zee, RNA folding and large n matrix theory, Nuclear Physics B 620 (2002) 456–476.

[18] C. M. Reidys, F. Huang, J. E. Andersen, R. C. Penner, P. F. Stadler, M. E. Nebel, Topology and prediction of RNA pseudoknots, Bioinformatics 27 (2011) 1076–1085.

[19] C. M. Reidys, R. R. Wang, A. Y. Zhao, Modular, k-noncrossing diagrams, the electronic journal of combinatorics 17 (R76) (2010) 1.

[20] M. Bon, G. Vernizzi, H. Orland, A. Zee, Topological classification of RNA structures, J. Mol. Biol. 379 (2008) 900–911.

[21] R. Penner, M. S. Waterman, Spaces of RNA secondary structures, Advances in Mathematics 101 (1) (1993) 31–49.

[22] R. C. Penner, Cell decomposition and compactification of Riemann's moduli space in decorated Teichmüller theory, in: N. Tongring, R. C. Penner (Eds.), Woods Hole Mathematics-perspectives in math and physics, World Scientific, Singapore, 2004, pp. 263–301, arXiv: math.GT/0306190.

[23] R. C. Penner, M. Knudsen, C. Wiuf, J. E. Andersen, Fatgraph models of proteins, Comm. Pure Appl. Math. 63 (2010) 1249–1297.

[24] J. E. Andersen, F. W. Huang, R. C. Penner, C. M. Reidys, Topology of RNA-RNA interaction structures, Journal of Computational Biology 19 (7) (2012) 928–943.

[25] G. Chapuy, The structure of unicellular maps, and a connection between maps of positive genus and planar labelled trees, Probability Theory and Related Fields 147 (3) (2010) 415–447.

[26] G. Chapuy, A new combinatorial identity for unicellular maps, via a direct bijective approach, Adv. Appl. Math. 47(4) (2011) 874–893.

[27] F. W. Huang, M. E. Nebel, C. M. Reidys, Generation of RNA pseudoknot structures with topological genus filtration, Mathematical biosciences 245 (2013) 216–225.

[28] F. W. Huang, C. M. Reidys, Shapes of topological RNA structures, arXiv:1403.2908.

[29] A. S. Richter, R. Backofen, Accessibility and conservation: General features of bacterial small RNA–mRNA interactions?, RNA Biology 9 (2012) 954–965.

[30] P. Duchon, P. Flajolet, G. Louchard, G. Schaeffer, Boltzmann samplers for the random generation of combinatorial structures, Comb. Probab. Comput. 13 (4-5) (2004) 577–625.

[31] D. H. Mathews, J. Sabina, M. Zuker, D. H. Turner, Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure, J. Mol. Biol. 288 (1999) 911–940.

[32] M. S. Waterman, Combinatorics of RNA hairpins and cloverleaves, Studies Appl. Math 60 (1978) 91–96.

[33] M. S. Waterman, Secondary structure of single–stranded nucleic acids, Adv. math. suppl. studies 1 (1978) 167–212.

[34] M. Zuker, D. Sankoff, RNA secondary structures and their prediction, Bulletin of mathematical biology 46 (4) (1984) 591–621.

[35] R. Nussinov, G. Pieczenik, J. R. Griggs, D. J. Kleitman, Algorithms for loop matchings, SIAM Journal on Applied mathematics 35 (1) (1978) 68–82.

[36] D. Kleitman, Proportions of irreducible diagrams, Studies in Applied Mathematics 49 (3) (1970) 297.

[37] H. S. Han, C. Reidys, A bijection between unicellular and bicellular maps, arXiv:1301.7177.